AFRL-RH-WP-TR-2010-0067

Speech Recognition, Articulatory Feature Detection, and Speech Synthesis in Multiple Languages

Brian M. Ore

General Dynamics Advanced Information Systems
5200 Springfield Street, Suite 200
Dayton OH 45431-1265

November 2009

Interim Report for August 2004 to November 2009

Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Anticipate & Influence Behavior Division
Sensemaking & Organizational Effectiveness Branch
Wright-Patterson AFB OH 45433-7022

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2010-0067 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNED//
RAYMOND E. SLYH
Work Unit Manager
Sensemaking & Organizational
Effectiveness Branch

//SIGNED//
GLENN W. HARSHBERGER
Anticipate & Influence Behavior Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* November 2009	**2. REPORT TYPE** Interim	**3. DATES COVERED** *(From - To)* August 2004 – November 2009

4. TITLE AND SUBTITLE Speech Recognition, Articulatory Feature Detection, and Speech Synthesis in Multiple Languages	**5a. CONTRACT NUMBER** FA8650-04-C-6443
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 62202F
6. AUTHOR(S) Brian M. Ore	**5d. PROJECT NUMBER** 7184
	5e. TASK NUMBER 08
	5f. WORK UNIT NUMBER 71840871
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) General Dynamics Advanced Information Systems 5200 Springfield Street, Suite 200 Dayton OH 45431-1265	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command Air Force Research Laboratory 711th Human Performance Wing Human Effectiveness Directorate Anticipate & Influence Behavior Division Sensemaking & Organizational Effectiveness Branch Wright-Patterson AFB OH 45433-7022	**10. SPONSOR/MONITOR'S ACRONYM(S)** 711 HPW/RHXS
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RH-WP-TR-2010-0067

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
88ABW cleared on 30 April 2010, 88ABW-2010-2329.

14. ABSTRACT

This document provides a summary of work completed by General Dynamics under the work unit 71840871, Speech Interfaces for Multinational Collaboration, for the period August 2004 to February 2009 under contract FA8650-04-C-6443. The speech technologies developed during this period include speech recognizers, Articulatory Feature (AF) detectors, and speech synthesizers. Speech recognition systems were developed for 15 different languages, and three methods were investigated for improving the performance of the systems: vocal tract length normalization, speaker adaptive training, and recognizer output voting error reduction. English AF detectors were developed using Gaussian mixture models, two-class Multi-Layer Perceptrons (MLPs), fusion MLPs, and multi-class MLPs. The outputs of the AF detectors were used to form the feature set for a speech recognizer. Speech synthesis systems were created for 13 different languages, and the following system modifications were investigated: expanding the label set to include additional contextual factors, changing the minimum description length control factor, and applying speaker clustering and adaption to create new voices. In addition, two graphical user interfaces were developed for training new voices and synthesizing speech in real-time.

15. SUBJECT TERMS Speech Recognition, Speech Synthesis, Articulatory Detection

16. SECURITY CLASSIFICATION OF: Unclassified			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Raymond E. Slyh
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	36	**19b. TELEPHONE NUMBER** *(include area code)* NA

THIS PAGE LEFT INTENTIONALLY BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE LEFT INTENTIONALLY BLANK

SUMMARY

This document provides a summary of work completed by General Dynamics under the work unit 71840871, Speech Interfaces for Multinational Collaboration, for the period August 2004 to November 2009 under contract FA8650-04-C-6443. The speech technologies developed during this period include speech recognizers, Articulatory Feature (AF) detectors, and speech synthesizers. Speech recognition systems were developed for 15 different languages, and three methods were investigated for improving the performance of the systems: vocal tract length normalization, speaker adaptive training, and recognizer output voting error reduction. English AF detectors were developed using Gaussian Mixture Models (GMMs), two-class Multi-Layer Perceptrons (MLPs), fusion MLPs, and multi-class MLPs. The outputs of the AF detectors were used to form the feature set for a speech recognizer. Speech synthesis systems were created for 13 different languages, and the following system modifications were investigated: expanding the label set to include additional contextual factors, changing the minimum description length control factor, and applying speaker clustering and adaption to create new voices. In addition, two graphical user interfaces were developed for training new voices and synthesizing speech in real-time.

The author would like to acknowledge the following groups: (1) Army Research Laboratory for the Dari speech corpus, (2) Cambridge University for their Hidden Markov Model ToolKit (HTK), (3) Carnegie Mellon University and Cambridge University for their Statistical Language Modeling Toolkit, (4) the Julius project team at Nagoya Institute of Technology for their Julius speech recognition engine, (5) Bryan Pellom of the University of Colorado for the SONIC speech recognizer, (6) the speech processing group at the Brno University of Technology, Faculty of Information Technology for their SRover software, (7) MIT Lincoln Laboratory for their GMM software package, (8) the International Computer Science Institute for their QuickNet software package, (9) Joe Frankel et al. for their AF classifiers, (10) Simon King et al. for the SVitchboard I corpus, (11) Karen Livescu et al. for their manual AF transcriptions, (12) the HTS working group for their HMM-based Speech Synthesis Toolkit, (13) the SPTK working group at Nagoya Institute of Technology for their Speech Signal Processing ToolKit, and (14) KTH for the Snack Sound Toolkit.

1.0 INTRODUCTION

This document provides a summary of work completed by General Dynamics under the work unit 71840871, Speech Interfaces for Multinational Collaboration, for the period August 2004 to November 2009 under contract FA8650-04-C-6443. The Section 2 describes how speech recognition systems were developed for 15 different languages, and presents three methods that were investigated for improving the performance of these systems. Section 3 describes how articulatory feature detectors were created for English and applied to speech recognition tasks in English, Russian, and Dari. Section 4 describes how speech synthesis systems were developed for 13 different languages, and provides a brief overview of two graphical user interfaces that were developed for creating new voices and synthesizing speech. Finally, Section 5 summarizes the work completed and provides recommendations for future research.

2.0 SPEECH RECOGNITION IN 15 LANGAUGES

Speech recognition systems were developed for 15 different languages using the Hidden Markov Model (HMM) ToolKit (HTK). This chapter discusses these recognition systems and presents three methods that were investigated to improve the performance of these systems: Vocal Tract Length Normalization (VTLN), Speaker Adaptive Training (SAT), and the ROVER technique. Section 2.1 provides an overview of the baseline recognition systems developed for each language. Section 2.2 discusses VTLN and presents results obtained on English, Mandarin, and Russian. Section 2.3 provides an overview of SAT and presents results obtained on Russian and Dari. Lastly, Section 2.4 describes the ROVER technique.

2.1 Baseline Recognition Systems

This section discusses the baseline speech recognition systems that were developed for Arabic, Croatian, Dari, English, French, German, Japanese, Korean, Mandarin, Pashto, Russian, Spanish, Tagalog, Turkish, and Urdu. A total of seven different corpora were used to obtain coverage of all 15 languages, including the Topic Detection and Tracking (TDT4) Multilingual Broadcast News corpus [1], Phase II of the Wall Street Journal (WSJ1) corpus [2], CALLHOME Mandarin Chinese [3], HUB4 Mandarin Broadcast News Speech [4], GlobalPhone [5], the Language And Speech Exploitation Resources (LASER) Advanced Concept Technology Demonstration corpus, and the ARL Dari corpus. The TDT4, WSJ1, CALLHOME, and HUB4 corpora are available from the Linguistic Data Consortium, and the ARL Dari corpus was collected by Army Research Laboratory with support from AFRL. Table 1 lists the corpora used for each language, the speaking style of each corpus, the total amount of training data used to develop the recognizers, and the vocabulary size.

Table 1: Overview of Corpora

Language	Corpus	Speaking Style	Hours	Vocabulary Size
Arabic	TDT4	Broadcast News	37	47k
Croatian	GlobalPhone	Read	12	22k
Dari	ARL	Read	20	2k
English	WSJ1	Read	18	10k
French	GlobalPhone	Read	20	21k
German	GlobalPhone	Read	14	23k
Japanese	GlobalPhone	Read	26	18k
Korean	GlobalPhone	Read	16	50k
Mandarin	CALLHOME	Conversational	26	8k
Mandarin	HUB4	Broadcast News	30	18k
Pashto	LASER	Read	17	6k
Russian	GlobalPhone	Read	18	29k
Spanish	GlobalPhone	Read	17	19k
Tagalog	LASER	Read	9	5k
Turkish	GlobalPhone	Read	13	15k
Urdu	LASER	Read	45	8k

HMM-based recognition systems were trained for each language using HTK [6].[1] The feature set consisted of 12 Mel-Frequency Cepstral Coefficients (MFCCs), with cepstral mean subtraction, plus an energy feature. Delta and acceleration coefficients were also included to form a 39 dimensional feature set. The acoustic models were state-clustered cross-word triphones. All HMMs included three states, with diagonal covariance matrices, and the state clustering was performed using a decision tree. An average of 16 mixture components were used for each HMM state.

Trigram Language Models (LMs) were created for each language using the Carnegie Mellon University (CMU)-Cambridge Toolkit [7].[2] The LM probabilities were estimated using the train partition of each language, but the vocabulary was expanded to include all words in the corpus. Decoding was performed using both the HTK decoder HDecode and the Julius decoder [8].[3] The Word Error Rates (WERs) for each language are shown in Figure 1. HDecode yielded better performance than Julius in all languages.

[1] Available at http://htk.eng.cam.ac.uk

[2] Available at http://www.speech.cs.cmu.edu

[3] Available at http://julius.sourceforge.jp

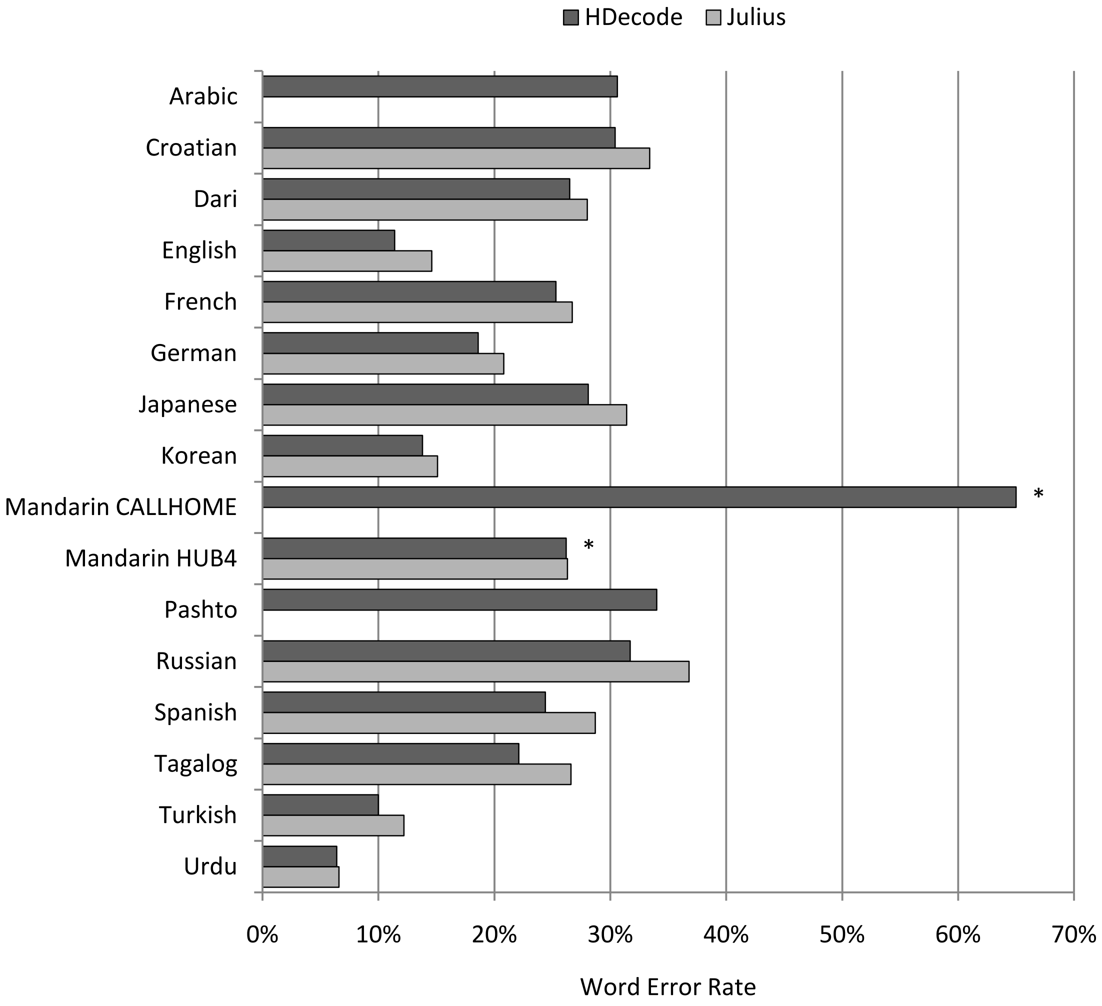


Figure 1: WER for each Language (HDecode and Julius); (*Mandarin is expressed in character error rate)

2.2 Vocal Tract Length Normalization

Vocal Tract Length Normalization (VTLN) attempts to compensate for different vocal tract lengths by linearly warping the frequency axis when performing filterbank analysis. Warping factors α for each speaker in the training set were selected using the following procedure [9]. First, single-mixture monophone HMMs with non-normalized MFCC features[4] were estimated from the complete training set of all speakers. Next, each utterance was phonemically aligned using the non-normalized HMMs and MFCC features computed using warping factors α=0.80,0.82,0.84,...,1.20. The value of α that gave the maximum score was selected for each speaker. Lastly, multiple-mixture triphone HMMs were estimated from the complete training set using the normalized MFCC features.

The procedure used to select the warping factor α for each utterance in the test set can be

[4] The term *normalization* is used to here to refer to MFCC features computed from a warped filterbank using α

summarized as follows. First, non-normalized multiple-mixture triphone HMMs with non-normalized MFCC features were used to hypothesize the word sequence for the utterance. Next, the utterance was phonemically aligned using the normalized single-mixture monophone HMMs and MFCC features computed using warping factors α=0.80,0.82,0.84,...,1.20. The value of α that gave the maximum score was selected for the utterance. Lastly, the normalized multiple-mixture triphone HMMs and normalized MFCC features were used to hypothesize the word sequence. The VTLN procedure was evaluated on the WSJ1 English, CALLHOME Mandarin, and GlobalPhone Russian. The results for each language are shown in Table 2. Applying VTLN reduced the error rate by 1.0% on English, 1.7% on Mandarin, and 0.3% on Russian.

Table 2: WER for English and Russian, and Character Error Rate for Mandarin

Language	No VTLN	With VTLN
English	11.8%	10.8%
Mandarin	65.1%	63.4%
Russian	29.6%	29.3%

2.3 Speaker Adaptive Training

Speaker Adaptive Training (SAT) is a technique used to train Speaker Independent (SI) acoustic models that integrates speaker normalization as part of the model estimation procedure. The procedure used to implement SAT can be summarized as follows. First, multiple-mixture triphone HMMs were estimated from the complete training set of all speakers. Next, Constrained Maximum Likelihood Linear Regression (CMLLR)[5] was used to compute a set of linear transformations for each speaker. Lastly, the SI models were re-estimated using the speaker transforms to adapt the training features. This procedure was repeated three times to train the final model.

The decoding procedure can be summarized as follows. First, the original SI acoustic models were used to hypothesize the word sequence for each utterance. Next, each utterance was phonemically aligned using the SI acoustic models. These alignments were used to compute a single set of CMLLR transforms for each speaker using the SAT models. Lastly, the SAT models and CMLLR transforms were used to hypothesize the word sequence for each utterance. The SAT technique was evaluated on the GlobalPhone Russian and ARL Dari. The results are shown in Table 3. Applying SAT reduced the WER by 4.5% on Russian and 3.1% on Dari.

Table 3: SAT for Russian and Dari

Language	No SAT	With SAT
Russian	29.6%	25.1%
Dari	26.6%	23.5%

[5] CMLLR is a feature adaptation technique that shifts the feature vectors such that each HMM state in the model is more likely to have generated the features

2.4 ROVER

Recognizer Output Voting Error Reduction (ROVER) [10] is a technique for combining the hypothesized word sequences from multiple recognizers. The ROVER technique first aligns the word sequences output from the different recognizers and then selects the final word sequence according to the frequency of occurrence. This technique was evaluated on 12 different languages using the hypothesized word sequences from the HDecode, Julius, and SONIC [11] decoders. The SRover program from the Brno University of Technology[6] was used to apply ROVER. Figure 2 shows the error rates obtained on each language. An improvement in system performance was obtained on all languages except English. Compared to the best individual system, the largest decrease in WER was 2.4% on French.

[6] Available at http://speech.fit.vutbr.cz

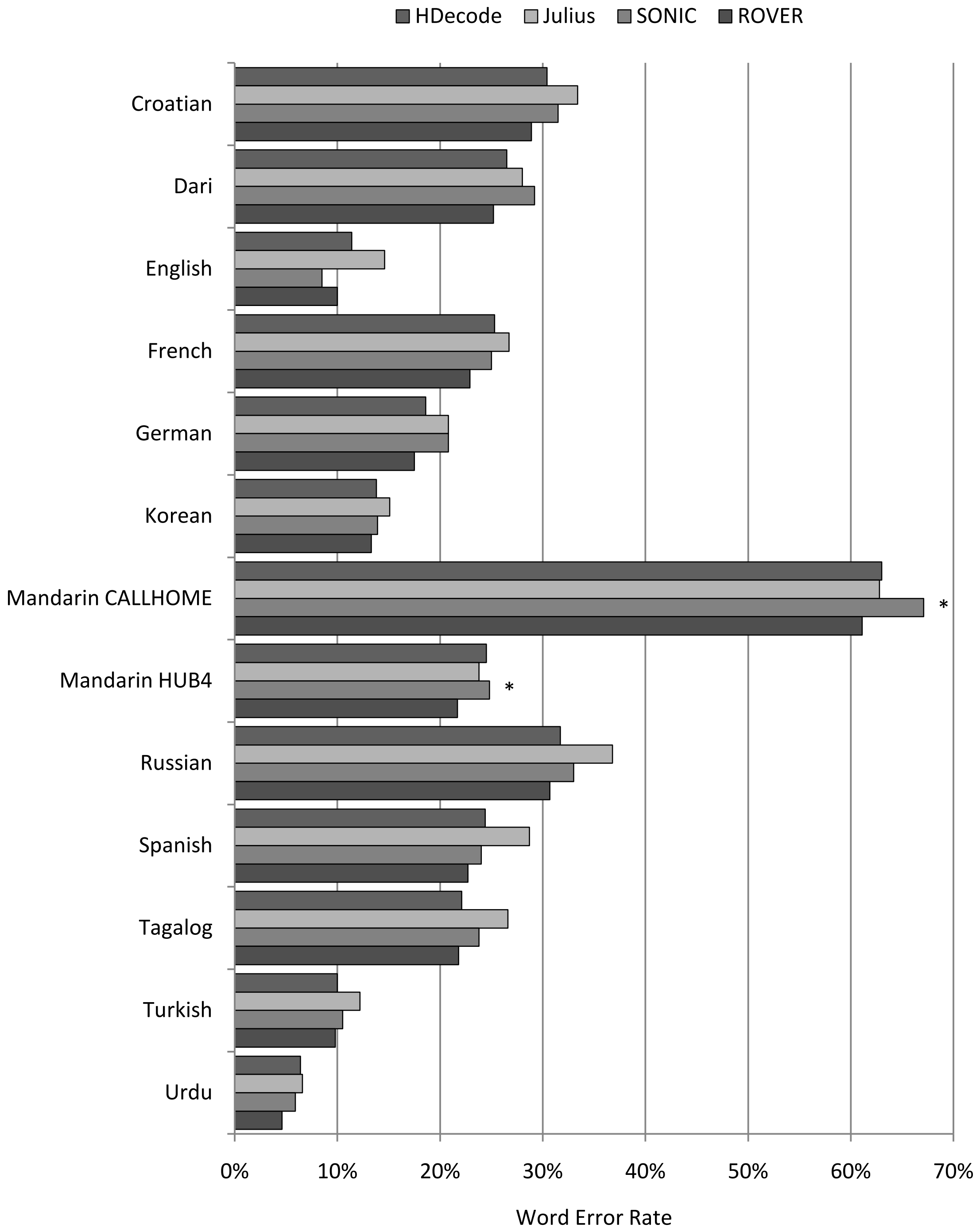


Figure 2: WER for each Language (HDecode, Julius, SONIC and ROVER); (*Mandarin is expressed in character error rate)

3.0 ARTICULATORY FEATURE DETECTION

Articulatory Features (AFs) describe the way in which speech sounds are produced. One of the most popular methods for classifying speech sounds using AFs is the International Phonetic Alphabet (IPA) [12]. Consonants are defined by AFs that describe the place of articulation, manner of articulation, and voicing status. Vowels are classified using AFs that describe both the tongue position and the shape of the lips. This chapter discusses two methods that were investigated for detecting English AFs. Section 3.1 describes how fusion-based AF detectors were created using Gaussian Mixture Models (GMMs) and two-class Multi-Layer Perceptrons (MLPs). Section 3.2 describes how multi-class MLPs were developed for English and incorporated into a Russian and Dari speech recognizer.

3.1 Fusion-based AF Detectors

This section discusses how fusion-based AF detectors were created for English and used in an HMM-based phoneme recognizer. Sections 3.1.1 and 3.1.2 describe how GMMs and MLPs were used to create AF detectors. Section 3.1.3 discusses two different procedures that were investigated for fusing the scores from the GMMs and MLPs, and presents results obtained on TIMIT. Lastly, Section 3.1.4 presents results obtained on the CSLU Multi-language Telephone corpus. Table 4 lists the AFs used to describe English speech sounds, with the exception of *silence (34)*, where the number in parenthesis indicates the feature number.

Table 4: AFs for English Consonants and Vowels

CONSONANTS (0)	
Place	bilabial (1), labiodental (2), labialvelar (3), dental (4), alveolar (5), postalveolar (6), retroflex (7), palatal (8), velar (9), glottal (10)
Manner	plosive (11), nasal (12), tap or flap (13), fricative (14), approximant (15), lateral approximant (16), affricate (17)
Voicing	voiced (18), voiceless (19)

VOWELS (20)	
Tongue Height	close (21), near-close (22), mid (23), open-mid (24), near-open (25), open (26)
Tongue Fronting	front (27), near-front (28), central (29), near-back (30), back (31)
Lip Shape	rounded (32), unrounded (33)

3.1.1 GMM-based AF Detectors. GMM-based AF detectors were trained on the WSJ1 corpus using the GMM software package from MIT Lincoln Laboratory [13]. For each AF, a GMM was trained using frames where the feature was present, and a second GMM was trained using frames where the feature was absent. All models used 256 mixture components with diagonal covariance matrices. The feature set consisted of 12 MFCCs, with cepstral mean subtraction, plus an energy feature. Delta and acceleration coefficients were also included to form a 39 dimensional feature vector.

The scores for each AF were calculated as follows. Denote the presence of an AF as *f* and the absence of an AF as *g*. If we consider the speech feature vector *x*, then

$$\log \frac{p(f \mid x)}{p(g \mid x)} = \log p(x \mid f) - \log p(x \mid g) + \log p(f) - \log p(g) \qquad (1)$$

The probabilities *p(x|f)* and *p(x|g)* were calculated from the feature-present and feature-absent GMMs, respectively. The probabilities *p(f)* and *p(g)* were estimated from the training data by counting the occurrences of each AF.

3.1.2 MLP-based AF Detectors. MLP-based AF detectors were trained on the WSJ1 corpus using the ICSI QuickNet software package.[7] A three-layered MLP (input: 39 units, hidden: 100 units, output: 2 units) was used to model each AF. The same MFCC feature set described in Section 3.1.1 was used as the input, and sigmoid activation functions were used on the hidden layer. The softmax function was used as the output activation function during training; however, it was removed when scoring the MLPs so that the outputs more closely approximated a Gaussian distribution. The final score for each AF was calculated by subtracting the output of the absent unit from the output of the present unit.

3.1.3 Score Fusion on TIMIT. This section describes two procedures that were investigated for fusing the scores from the GMM- and MLP-based AF detectors [14]. Both methods trained a fusion MLP for each AF to combine the scores. All fusion MLPs were trained on the TIMIT corpus [15]. *Fusion-1* combined the scores from the GMM- and MLP-based AF detectors for a given AF to form the final score for that AF. For example, the fusion MLP for the AF *plosive* used input features consisting of the output of the GMM-based plosive detector and the MLP-based plosive detector. *Fusion-2* combined the scores from all of the GMM- and MLP-based AF detectors to form the final score for each AF; thus, the fusion MLP for each AF was provided information about all AFs from two different classifiers.

All fusion MLPs included 100 hidden units with sigmoid activation functions, and used the softmax output activation function for training. The fusion MLPs included a context window of nine; that is, the MLPs used the vectors at times t-4,t-3,$\cdots$,t+3,t+4 as input to classify the vector at time *t*. As in Section 3.1.2, the output activation function was removed prior to scoring and the score for the AF was calculated by subtracting the output of the absent unit from the output of the present unit.

Figure 3 shows the AF detection results obtained on the TIMIT test set. Each symbol represents the average Equal Error Rate (EER) of the individual detectors for the AF groups shown in Table 4. For the place and manner classifiers, the GMM-based detectors outperformed

[7] Available at http://www.icsi.berkeley.edu/Speech/qn.html

the MLP-based detectors; for all other groups the MLPs yield lower EERs. Fusion-1 yielded an average decrease in EER of 4.7% absolute compared to the best GMM- or MLP-based detector.[8] The best overall performance was obtained using the Fusion-2 procedure, which yielded an average decrease in EER of 8.2% absolute compared to the best GMM- or MLP-based detector.

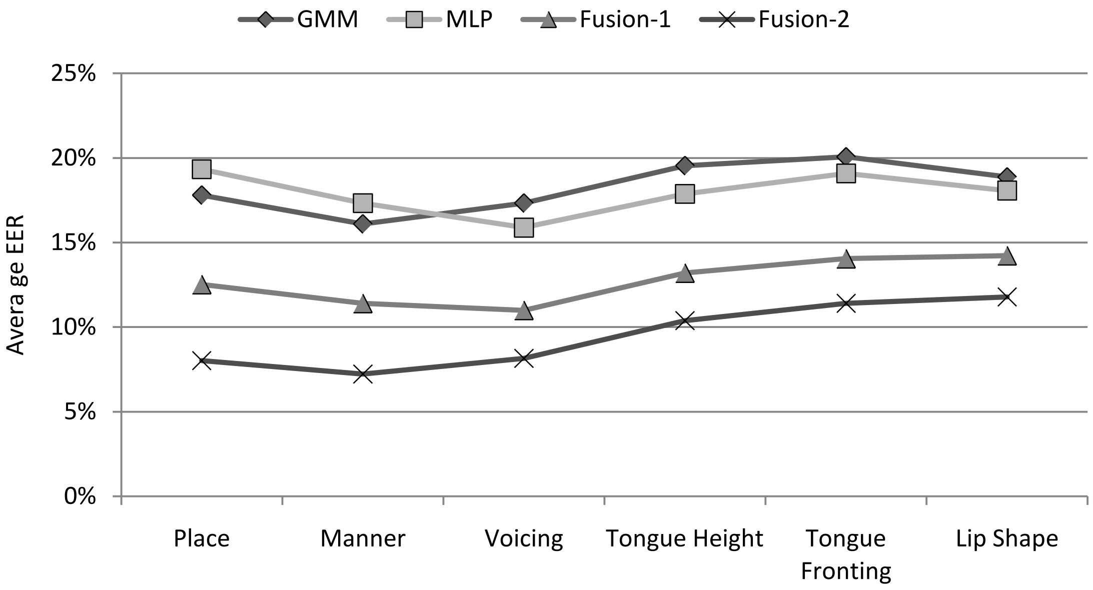


Figure 3: Average EER of the AF Detectors on the TIMIT Test Set

The scores from the different AF detectors were used to form the feature set for an HMM-based phoneme recognizer. First, a vector was formed using the scores from the individual AF detectors. Next, these feature vectors were processed with a Karhunen-Loéve Transformation (KLT) that was estimated on the TIMIT train set. The KLT was included to decorrelate the individual AF scores so that diagonal covariance matrices could be used in the HMMs. Lastly, delta features were appended. Monophone and triphone HMMs were created for each feature set. All systems used three state HMMs with 16 mixtures per state and diagonal covariance matrices. Decoding was performed using a bigram phoneme LM that was estimated from the TIMIT train set using the CMU-Cambridge Toolkit. The MFCC feature set described in Section 3.1.1 was used for the baseline system.

Table 5 shows the Phoneme Error Rate (PER) obtained with each feature set on the TIMIT test set. The features created using the scores from the GMM-based detectors yielded the worst performance. An improvement in recognition performance was obtained using the scores from the MLP-based detectors, however, the PER was still higher than that of the baseline MFCC system. The Fusion-1 features outperformed both the GMM and MLP features sets, although an increase in performance over the baseline MFCC system was only obtained with monophone models. The best performance was obtained using the Fusion-2 features.

[8] The term *best* is used here to refer to the detector with the minimum EER for each AF

Table 5: PER Obtained on the TIMIT Test Set

	MFCC	GMM	MLP	Fusion-1	Fusion-2
Monophones	39.5%	42.1%	39.9%	38.8%	35.8%
Triphones	35.9%	40.8%	38.4%	38.4%	35.6%

It is worth noting that the Fusion-2 monophone system yielded comparable performance to the MFCC triphone system. The option of using monophone instead of triphone models with the Fusion-2 features can be a significant advantage in terms of decoding time. Excluding the time required for feature extraction, decoding with each triphone system took approximately 750 minutes, whereas decoding with monophones was completed in about 20 minutes.

3.1.4 Score Fusion on CSLU. This section discusses AF detection on the CSLU Multi-Language corpus [16]. Whereas TIMIT consists of lab-quality recordings of read speech with broad phonetic coverage, the CSLU corpus includes spontaneous telephone speech. Thus, these corpora differ in speaking style (read vs. spontaneous), channel type (close-talking microphone vs. telephone), balance of phonetic coverage, and sampling rate.

The WSJ1l and TIMIT corpora were first downsampled to 8 kHz and a second set of Fusion-2 AF detectors were retrained. Next, a set of Fusion-2 AF detectors were trained on the CSLU corpus. All AF detectors were created using the same procedure described in Sections 3.1.1-3.1.3. It should be emphasized that all fusion MLPs used scores from GMM- and MLP-based detectors trained on WSJ1 as input. Thus for the CSLU corpus, the base GMM- and MLP-based detectors were used for a different speaking style (read vs. spontaneous) and channel (close-talking microphone vs. telephone).

Figure 4 shows the EERs obtained with the Fusion-2 AF detectors. Each symbol type represents a different *train-test* combination. For example, TIMIT8-CSLU shows the detection performance obtained on the CSLU test set using Fusion-2 AF detectors trained on the TIMIT corpus downsampled to 8 kHz. The individual symbols represent the EER of each AF detector, where the feature numbers correspond to those given in Table 4. The best overall performance was obtained on the TIMIT8-TIMIT8 condition. The average EER across all AFs was 8.6%. When evaluated on the CSLU corpus, the fusion MLPs trained on TIMIT8 yielded an average EER of 14.1%, which is an increase of 5.5% compared to the results on TIMIT8. The average EER of the Fusion-2 AF detectors trained and evaluated on CSLU was 11.5%.

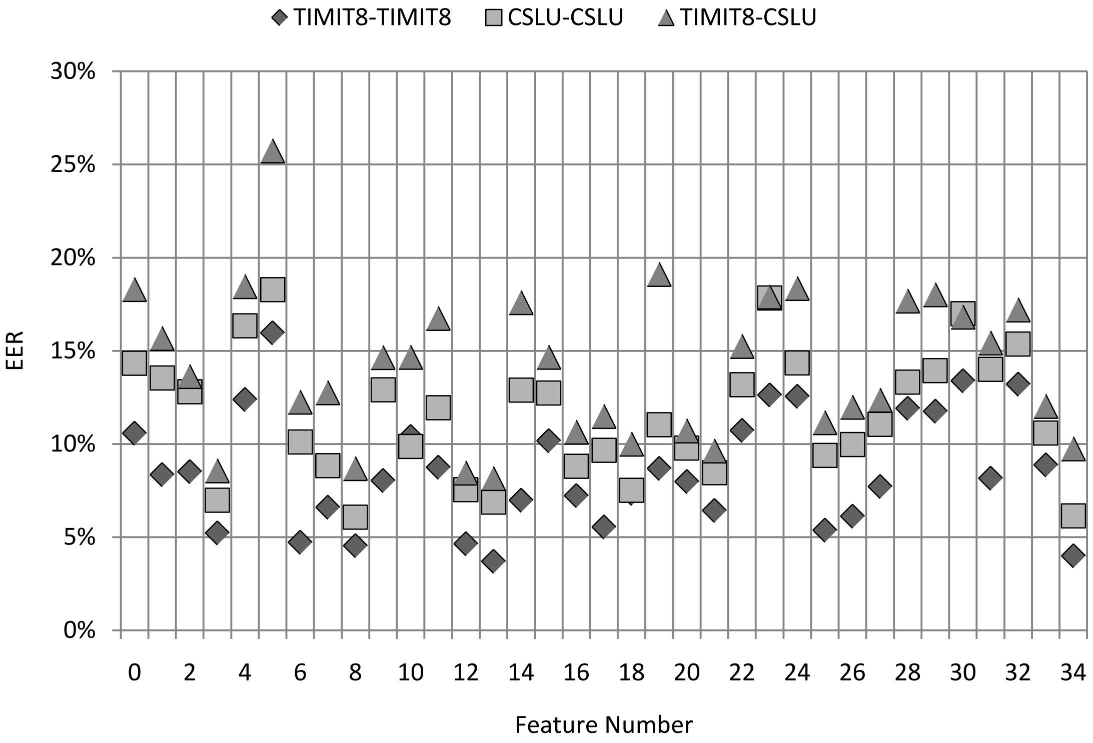


Figure 4: EER of the AF Detectors on the CSLU Test Set

From Figure 4 we can see that some of the AF detectors are more robust across both corpora than others. For example, the increase in EER on TIMIT8-CSLU compared to TIMIT8-TIMIT8 is less than 3.5% for the AFs *labialvelar (3), lateral approximant (16), voiced (18), vowel (20), close (21), near-back (30)*, and *unrounded (33)*. The increase in EER is greater than 8.0% for the AFs *alveolar (5), plosive (11), fricative (14)* and *voiceless (19)*. This suggests that certain AFs are less affected by speaking style and channel type than other AFs.

As in Section 3.1.3, the scores from the fusion MLPs were used to form the feature set for an HMM-based phoneme recognizer. Monophone and triphone HMMs were trained for each feature set on the CSLU corpus. The monophone models included 32 mixtures per state, and the triphone models included 12 mixtures per state. All systems used diagonal covariance matrices. Decoding was performed using a trigram phoneme LM that was estimated from the CSLU train partition using the CMU-Cambridge Toolkit. The MFCC feature set described in Section 3.1.1 was used for the baseline system. Table 6 shows the PER obtained with each feature set on the CSLU test set. Both the TIMIT8 and CSLU Fusion-2 feature sets outperform the MFCC system. The best performance was obtained with the CSLU Fusion-2 features: compared to MFCCs, the PER was reduced by 2.0% absolute when decoding with either monophone or triphone models.

Table 6: PER Obtained on the CSLU Test Set

	MFCC	TIMIT8 Fusion-2	CSLU Fusion-2
Monophones	49.4%	48.6%	47.4%
Triphones	48.3%	47.4%	46.3%

3.2 AF Detection using Multi-Class MLPs

This section discusses how multi-class MLPs were used to create English AF detectors. Section 3.2.1 describes the procedure used to train the MLPs. Section 3.2.2 presents detection results obtained on SVitchboard and describes how the scores from the MLPs were used as the feature set for a speech recognizer. Lastly, Section 3.2.3 presents results obtained on Russian and Dari. Table 7 lists the features that were used to describe English speech sounds.

Table 7: Features used to Describe English Speech Sounds [17]

Group	Feature Values
Place	alveolar, dental, labial, labiodental, lateral, none, postalveolar, rhotic, velar, silence
Degree	approximant, closure, flap, fricative, vowel, silence
Nasality	-, +, silence
Rounding	-, +, silence
Glottal State	aspirated, voiceless, voiced, silence
Vowel	aa, ae, ah, ao, aw1, aw2, ax, axr, ay1, ay2, eh, er, ey1, ey2, ih, iy, ix, ow1, ow2, oy1, oy2, uh, uw, none, silence
Height	high, low, mid, mid-high, mid-low, very-high, none, silence
Frontness	back, front, mid, mid-back, mid-front, none, silence

3.2.1 MLP-based AF Detectors. Two sets of MLPs were trained for each of the eight AF groups shown in Table 7. The first set used MFCCs as input, and the second set used Perceptual Linear Prediction (PLP) coefficients. The MFCC feature set was the same as described in section 3.1.1, except that both mean and variance normalization were applied on a per-conversation side basis. The PLP feature set included 12 PLP cepstral coefficients, plus energy, delta, and acceleration coefficients. As with the MFCCs, mean and variance normalization were also applied.

The MLPs were trained on the Fisher corpus [18, 19] using the ICSI QuickNet software package. A context window of nine was used on the input layer, and the number of hidden units for each MLP was chosen using the same procedure as described in [17]. Sigmoid activation functions were used on the hidden layer. The number of output units for each MLP was set to the number of feature values for that AF group, and the softmax function was used as the output activation function.

3.2.2 AF Detection on SVitchboard. This section discusses AF detection on the SVitchboard corpus [20]. SVitchboard is a small vocabulary corpus that includes conversational telephone speech. A subset of 78 utterances includes AF alignments that were manually produced [21]. Figure 5 shows the frame level accuracy of the MLPs trained on Fisher using MFCC and PLP coefficients as input. For comparison purposes, the detectors from [17] were also evaluated on these utterances. These detectors, referred to as *Frankel* in this document, use the same network typology and PLP feature set as the MLPs described in section 3.2.1. Overall, similar performance is obtained with each set of MLPs. The largest difference in accuracy is 2.0% (Frankel vs. PLP *degree*). The lowest accuracy was 75.8% (MFCC *place),* and the highest accuracy was 95.4% (Frankel *nasality).*

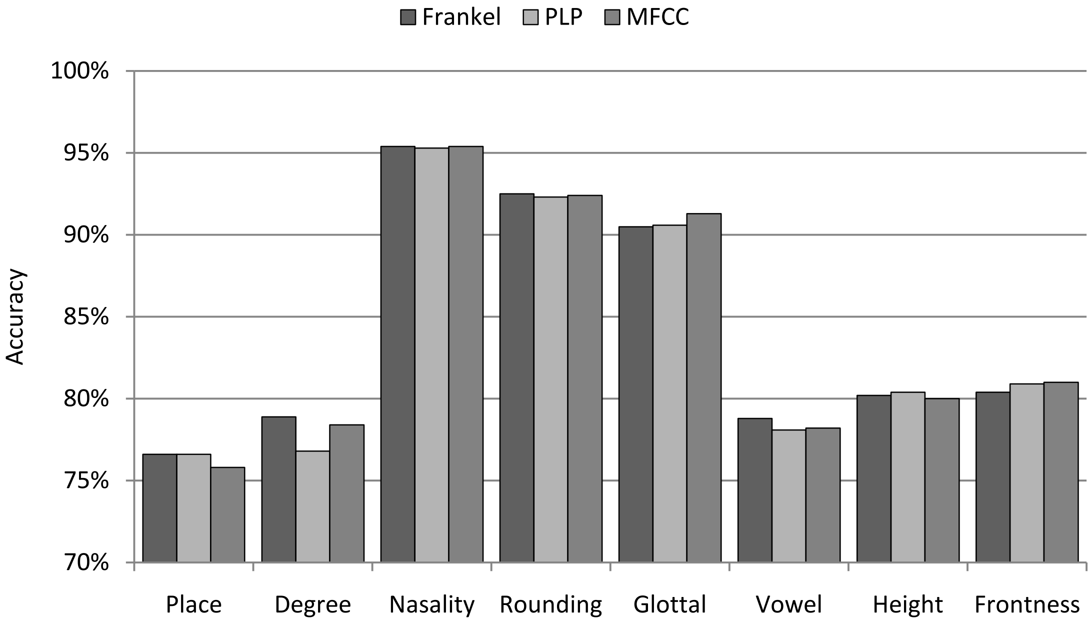


Figure 5: Frame Level Accuracy of the MLP-based AF Detectors on the SVitchboard Corpus

The scores from the MLPs were used to form the feature set for an HMM-based speech recognizer. First, a vector was formed using the scores from the individual AF detectors. When computing these scores, the output activation function was removed so that the scores more closely approximated a Gaussian. Next, these feature vectors were processed with a KLT that was estimated on the SVitchboard train set, and the top 26 dimensions were retained. This feature vector was appended to the PLP feature set described in Section 3.2.1 to form a 65 dimensional vector.

Within-word triphone HMMs were trained for each feature set. All systems used three state HMMs with 12 mixtures per state and diagonal covariance matrices. Decoding was performed using a bigram LM that was estimated from the SVitchboard train set using HTK.

The PLP features formed the baseline system. Table 8 shows the WER obtained with each system. From Table 8 we can see that incorporating the scores from the MLPs yielded an improvement in system performance. The best WER was obtained with the PLP system that incorporated the Frankel MLPs: compared to the baseline PLP system, a reduction in WER of 6.0% was obtained. Note also that the MLP system with PLP input features yielded better performance than the MLP system with MFCC input features.

Table 8: WER on the SVitchboard 500 Word Vocabulary Task

Features	WER
PLP	50.6%
PLP + Frankel	44.6%
PLP + MLPs with PLP input	44.8%
PLP + MLPs with MFCC input	46.0%

3.2.3 Cross-Lingual AF Detection. The Frankel MLPs were also evaluated on the GlobalPhone Russian and ARL Dari. Whereas the Frankel MLPs were trained on English conversational telephone speech, the GlobalPhone Russian and ARL Dari corpora consist of read microphone speech. Thus, these corpora differ not only in language, but also in speaking style (conversational vs. read), channel type (telephone vs. microphone), and sampling rate.

The GlobalPhone Russian and ARL Dari corpora were first downsampled to 8 kHz and PLP features were extracted. These features were used as input to the Frankel MLPs, which were evaluated with the output activation functions removed. Next, a vector was formed using the scores from the individual AF detectors and processed with a KLT that was estimated on the train partition of each language. The top 26 dimensions were retained and appended to the MFCC feature set described in Section 2.1. This feature vector was used to train an HMM-based speech recognizer for each language. The HMM systems were trained using the same procedure described in Section 2.1 and decoding was performed using HDecode. The WER for each language is shown in Table 9. Incorporating the Frankel MLPs reduced the WER by 1.6% on Russian and 1.4% on Dari.

Table 9: WER on Russian and Dari

Language	MFCC	MFCC + Frankel
Russian	29.6%	28.0%
Dari	26.4%	25.0%

4.0 SPEECH SYNTHESIS IN 13 LANGUAGES

Speech synthesis systems were developed for 13 different languages using the Hidden Markov Model (HMM) Speech Synthesis ToolKit (HTS). This chapter describes these systems and provides an overview of two different Graphical User Interfaces (GUIs) that were developed for creating new voices and synthesizing speech. Section 4.1 provides an overview of the baseline synthesis systems. Section 4.2 describes three English and two Urdu speech synthesis systems that were created using an expanded model set. Section 4.3 discusses the effect of modifying the Minimum Description Length (MDL) control factor. Section 4.4 discusses speaker clustering and adaptation for creating English and Mandarin voices. Lastly, Section 4.5 provides a brief overview of the GUIs that were developed.

4.1 Baseline Synthesis Systems

This section discusses the baseline synthesis systems that were developed for Arabic Iraqi, Croatian, Dari, English, French, German, Mandarin, Pashto, Russian, Spanish, Tagalog, Turkish, and Urdu. A total of six different corpora were used to obtain coverage of all languages, including the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) corpus, GlobalPhone, ARL, CMU Arctic [22], HUB4, and LASER. All of these corpora include speech data that were recorded with a 16 kHz sampling frequency. The CMU Arctic database was developed specifically for speech synthesis and includes automatically generated time-aligned transcriptions; all other corpora are only transcribed at the utterance level. Phoneme alignments for the TRANSTAC, GlobalPhone, ARL, HUB4, and LASER corpora were automatically generated using SONIC.

HMM-based speech synthesis systems were developed for each language using HTS-2.0 [23].[9] The feature set consisted of 25 Mel Cepstral Coefficients and the logarithm of the fundamental frequency (F0). Prior to computing the features, the DC mean was removed from each waveform file and amplitude normalization was applied to several of the corpora. The Mel Cepstral coefficients were calculated using the Speech Signal Processing ToolKit (SPTK),[10] and the F0 values were estimated using the ESPS method implemented in snack.[11] Delta and acceleration coefficients were also included to form a 78 dimensional feature vector.

Cross-word triphone Multi-Space probability Distribution (MSD)-HMMs [24] were trained for each language. All MSD-HMMs included five states with diagonal covariance matrices, and the state durations for each triphone were modeled by a Gaussian distribution. Decision tree based clustering was applied to the Mel Cepstrum, F0, and state duration distributions independently; thus, two decision trees were created for each MSD-HMM state, plus an additional decision tree for the state duration model. Table 10 lists the voices that were created for each language, the corpora used, the number of speakers used to train the voices, and the total amount of training data used to develop the synthesizers.

[9] Available at http://hts.sp.nitech.ac.jp

[10] Available at http://sp-tk.sourceforge.net

[11] Available at http://www.speech.kth.se/snack

Table 10: Overview of Voices Created

Language	Corpus	Voices	Speaker Count	Hours
Arabic Iraqi	TRANSTAC	Speaker1 Speaker2	370 30	10 3
Croatian	GlobalPhone	Male Female	32 48	5 7
Dari	ARL	Male1 Male2	15 15	2 2
English	CMU Arctic	Male Female SLT	4 2 1	3 2 1
French	GlobalPhone	Male Female	39 40	10 11
German	GlobalPhone	Male Female	60 5	13 1
Mandarin	HUB4	Male Wang Jianchuan Female Fang Jing	10 1 8 1	2 1 2 1
Mandarin	GlobalPhone	Male	15	4
Pashto	LASER	Random1 Random2	10 10	1 1
Russian	GlobalPhone	Male Female	49 44	9 9
Spanish	GlobalPhone	Male Female	38 46	8 10
Tagalog	LASER	Male Female	20 28	2 4
Turkish	GlobalPhone	Male Female	24 60	4 10
Urdu	LASER	Male Female	76 84	17 20

4.2 Full-Context Models

This section discusses the English and Urdu speech synthesizers that were created using an expanded model set. As mentioned in Section 4.1, the baseline synthesis systems for each language used cross-word triphone models. Although these models produce intelligible speech, there are numerous other contextual factors that can affect the overall prosody and naturalness of speech. In order to incorporate these contextual factors, the triphone labels for each speech database have to be expanded to include all features of interest. For example, the labels supplied with the HTS demos for the CMU Arctic database consist of 53 different contextual features, including syllable, accent, stress, part-of-speech, word, and phrase information. These labels are then used to define the acoustic models; thus, a separate MSD-HMM is trained for each phoneme that appears in a different context. Note that this can result in a very large model set prior to clustering. For example, the training data for the English SLT voice includes 38866 phoneme instances: using cross-word triphone labels requires 9480 unique MSD-HMMs, whereas using the expanded label set requires 38765 unique MSD-HMMs. An expanded set of labels were derived for Urdu that included syllable, word, and phrase information. These labels included a total of 31 different contextual features. Syllable information was explicitly marked in the pronunciation lexicon, and phrase information was derived by assigning a break wherever silence was labeled. Table 11 lists the expanded label set derived for Urdu.

Table 11: Expanded Label Set for Urdu

p1	the phoneme identity before the previous phoneme
p2	the previous phoneme identity
p3	the current phoneme identity
p4	the next phoneme identity
p5	the phoneme after the next phoneme identity
p6	position of the current phoneme in the current syllable (forward)
p7	position of the current phoneme in the current syllable (backward)
a1	the number of phonemes in the previous syllable
b1	the number of phonemes in the current syllable
b2	position of the current syllable in the current word (forward)
b3	position of the current syllable in the current word (backward)
b4	position of the current syllable in the current phrase (forward)
b5	position of the current syllable in the current phrase (backward)
b6	name of the vowel of the current syllable
c1	the number of phonemes in the next syllable
d1	the number of syllables in the previous word
e1	the number of syllables in the current word
e2	position of the current word in the current phrase (forward)
e3	position of the current word in the current phrase (backward)
f1	the number of syllables in the next word
g1	the number of syllables in the previous phrase
g2	the number of words in the previous phrase
h1	the number of syllables in the current phrase
h2	the number of words in the current phrase
h3	position of the current phrase in this utterance (forward)
h4	position of the current phrase in this utterance (backward)
i1	the number of syllables in the next phrase
i2	the number of words in the next phrase
j1	the number of syllables in this utterance
j2	the number of words in this utterance
j3	the number of phrases in this utterance

Each of the three English voices and the two Urdu voices were retrained using the expanded labels. Overall, there was not a substantial improvement in voice quality. This may be due to the limited amount of speech data available to train different models for each phoneme in a particular context.

4.3 MDL Control Factor

Decision tree clustering in HTS is based on the MDL criterion [25]. The MDL criterion is used for selecting the questions when splitting nodes, and deciding when to stop growing the decision trees. A control factor λ is used to weight the penalty that the MDL criterion imposes for model complexity. As λ is increased, the penalty for a large model become larger and the stopping criterion is met sooner (thus producing a decision tree with fewer leaves). The English male and female voices described in Section 4.2 were retrained using λ = 1.0,0.7,0.4. The total number of leaves obtained for each λ is shown in Figure 6. As λ is increased, the total number of leaves for each of the decision trees decreases.

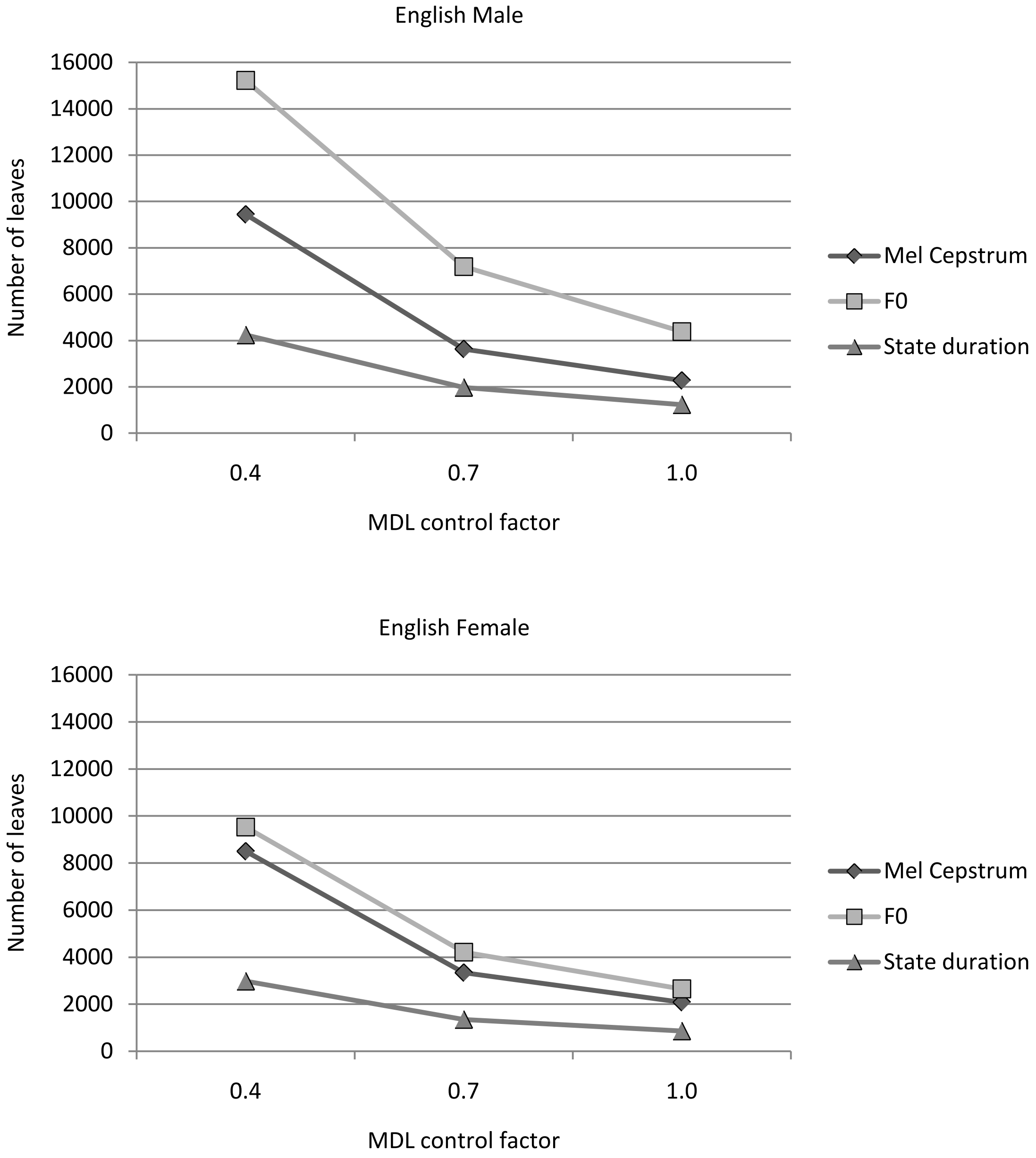


Figure 6: Total Number of Leaves Generated for the English Male and Female Voice when Modifying the MDL Control Factor λ

4.4 Speaker Clustering and Adaptation

This section discusses how speaker clustering and adaptation were used to create voices for Mandarin and English.[12] A total of 52 different Mandarin speech synthesis systems were trained on the GlobalPhone corpus using groups of three or more speakers. The speaker groups were defined based on the individual speakers F0 values and/or speaker recognition scores. Two additional voices were also created on the HUB4 Mandarin corpus by adapting the Male voice using speech from Wang Jianchuan, and adapting the Female voice using speech from Fang Jing. The adaptation transforms were estimated using Constrained Maximum Likelihood Linear Regression (CMLLR).

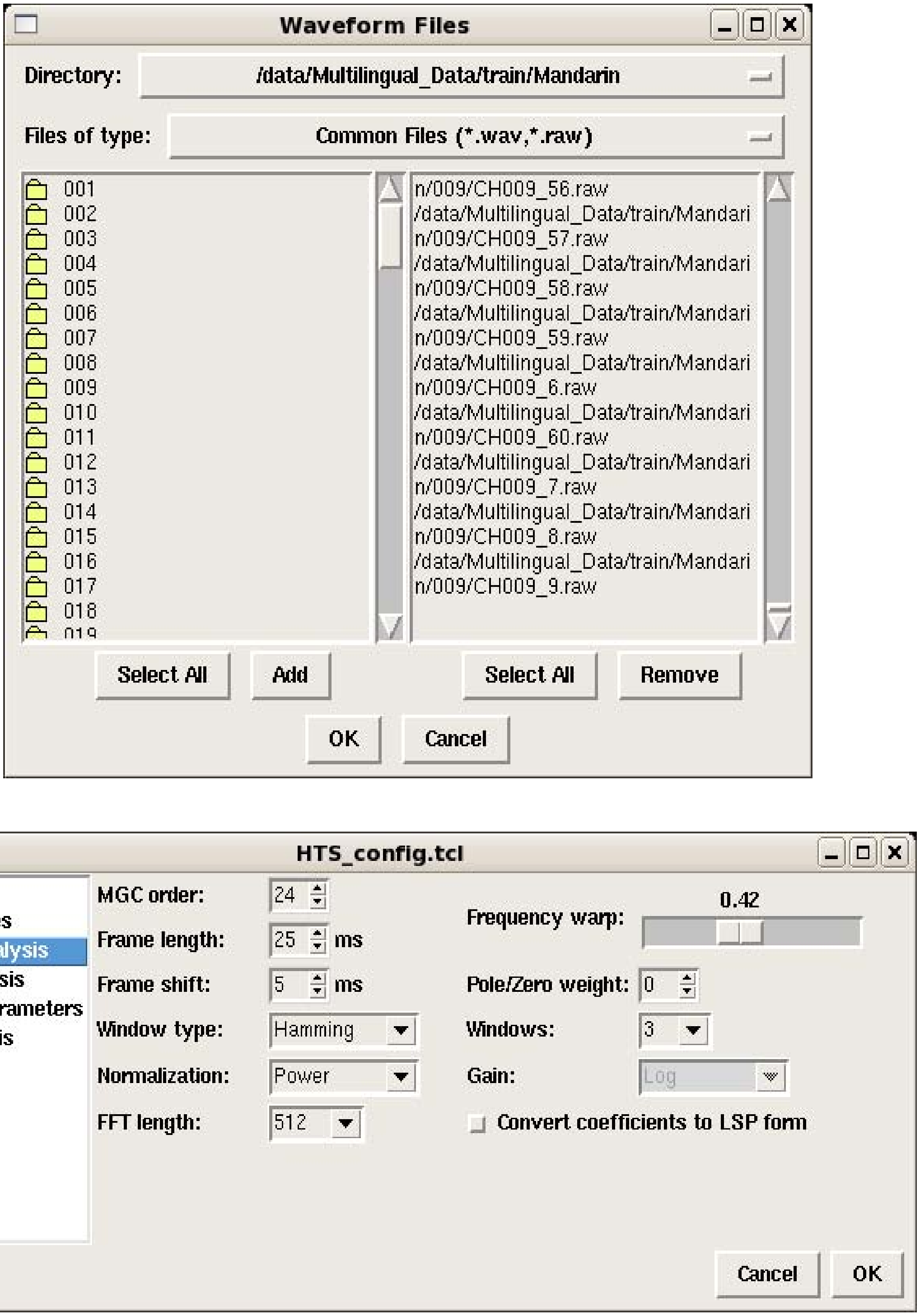
A total of 53 English speech synthesis systems were trained on Phase I of the Wall Street Journal (WSJ0) corpus [26] and WSJ1. These systems were developed using HTS-2.1.[13] Cross-word triphone MSD Hidden Semi-Markov Models (HSMMs) [27] were created for each voice using the same feature set as described in Section 4.1. As with the other corpora, the phoneme alignments were automatically generated using SONIC. The first 25 voices were created using groups of three or more speakers. The speaker groups were defined based on speaker recognition scores: 19 groups of speakers were derived from a speaker confusion matrix, and the remaining six groups were derived using a spectral clustering algorithm [28]. Next, one set of MSD-HSMMs were trained using 3600 utterances from nine different speakers (~400 utterances from each speaker), and a second set of MSD-HSMMs were trained using 3502 utterances from 20 different speakers (~200 utterances from each speaker). These models were adapted using speech from one of 22 different speakers to create the remaining 28 voices. Adaptation was performed using Constrained Structural Maximum-A-Posteriori Linear Regression (CSMAPLR), followed by MAP adaptation [29].

4.5 Synthesis GUIs

This section describes two GUIs that were developed for training and evaluating speech synthesizers. The first interface can be used to setup a speech synthesis experiment. This program allows the user to choose a set of speakers to train the voice and adjust system parameters related to speech analysis, model settings, and synthesis. Figure 7 shows two instances of the interface: the top one shows the speaker selection dialog, and the bottom one shows the spectrum analysis dialog. Once all configuration options have been specified, this program creates the makefiles for training and evaluating the system.

[12] The speaker recognition experiments, F0 analysis, and speaker cluster definitions described in this section (except for those derived using the spectral clustering algorithm) were generated by Mr. Eric Hansen

[13] Available at http://hts.sp.nitech.ac.jp



Waveform Files

Directory: /data/Multilingual_Data/train/Mandarin

Files of type: Common Files (*.wav,*.raw)

001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019

n/009/CH009_56.raw
/data/Multilingual_Data/train/Mandari
n/009/CH009_57.raw
/data/Multilingual_Data/train/Mandari
n/009/CH009_58.raw
/data/Multilingual_Data/train/Mandari
n/009/CH009_59.raw
/data/Multilingual_Data/train/Mandari
n/009/CH009_6.raw
/data/Multilingual_Data/train/Mandari
n/009/CH009_60.raw
/data/Multilingual_Data/train/Mandari
n/009/CH009_7.raw
/data/Multilingual_Data/train/Mandari
n/009/CH009_8.raw
/data/Multilingual_Data/train/Mandari
n/009/CH009_9.raw

HTS_config.tcl

MGC order: 24

Frame length: 25 ms

Frame shift: 5 ms

Window type: Hamming

Normalization: Power

FFT length: 512

Frequency warp: 0.42

Pole/Zero weight: 0

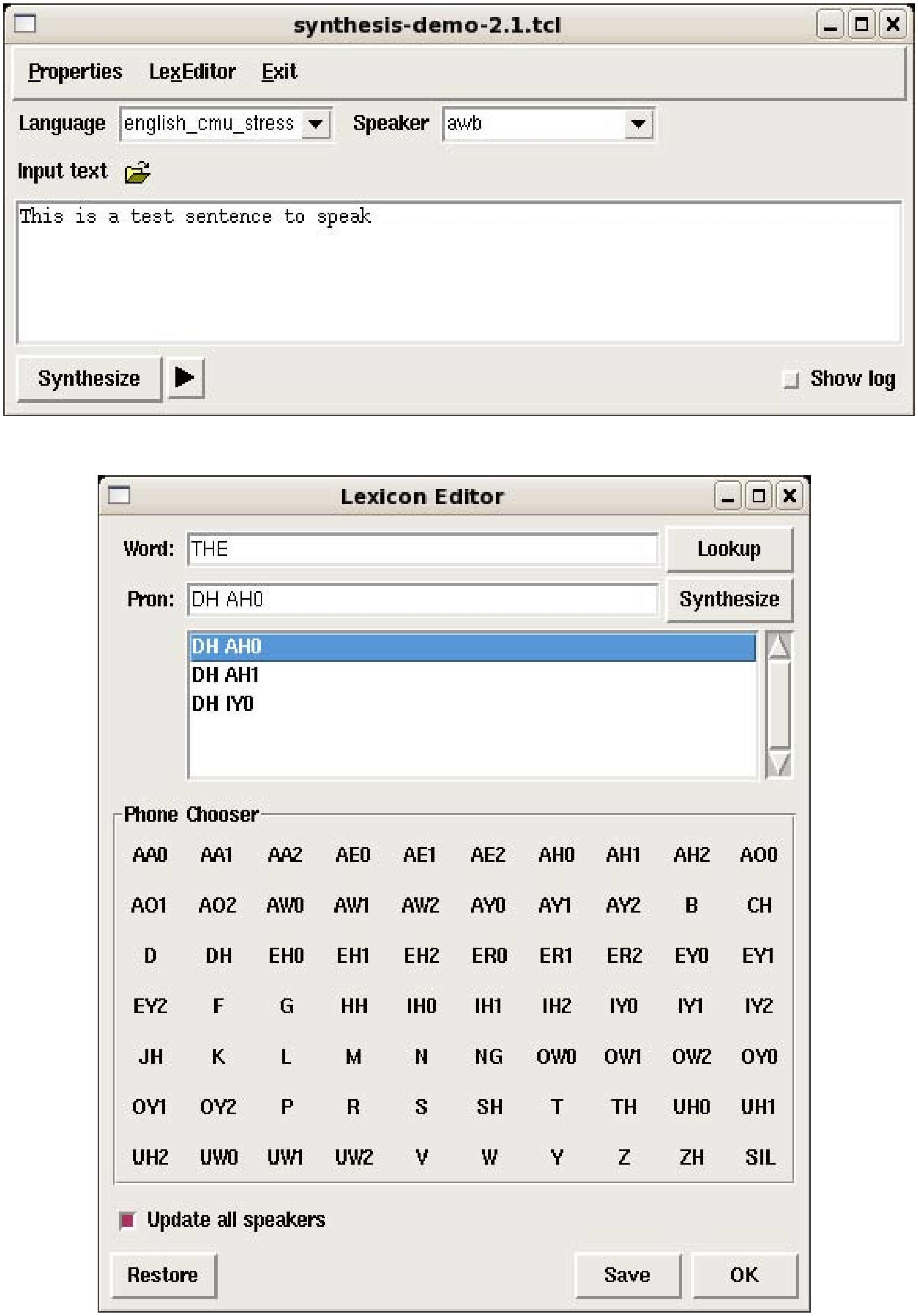
Windows: 3

Gain: Log

□ Convert coefficients to LSP form

Figure 7: GUI for Configuring a Speech Synthesis Experiment; speaker selection dialog is shown on top, and the spectrum analysis dialog is shown on the bottom

The second interface can be used to synthesize speech, modify pronunciations, and create new voices by modifying the synthesis parameters. The text to synthesize can be entered using either the keyboard or read from a text file, and the pronunciations can be modified and saved on a per-speaker basis. The following synthesis parameters can be modified: all-pass constant, post-filtering coefficient, speech speed rate, multiplicative and additive constants for F0, voiced/unvoiced threshold, spectrum and F0 global variance weights, amplitude normalization constant, maximum state duration variance, and model interpolation coefficients. Figure 8 shows the main interface and pronunciation editor.


synthesis-demo-2.1.tcl
Properties
LexEditor
Exit
Synthesize
Lexicon Editor
Lookup
Synthesize
Phone Chooser
AA0 AA1 AA2 AE0 AE1 AE2 AH0 AH1 AH2 AO0
AO1 AO2 AW0 AW1 AW2 AY0 AY1 AY2 B CH
D DH EH0 EH1 EH2 ER0 ER1 ER2 EY0 EY1
EY2 F G HH IH0 IH1 IH2 IY0 IY1 IY2
JH K L M N NG OW0 OW1 OW2 OY0
OY1 OY2 P R S SH T TH UH0 UH1
UH2 UW0 UW1 UW2 V W Y Z ZH SIL
Restore
Save
OK

Language english_cmu_stress   Speaker awb

Input text

```
This is a test sentence to speak
```

□ Show log

Word: THE

Pron: DH AH0

DH AH0
DH AH1
DH IY0

☑ Update all speakers

Figure 8: GUI for Synthesizing Speech; the main interface is shown on top, and the pronunciation editor is shown on the bottom

5.0 CONCLUSIONS AND RECOMMENDATIONS

This document summarized work completed by General Dynamics during the period August 2004 to February 2009. Speech recognition systems were developed for 15 different languages using HTK. Three methods were investigated for improving the performance of these systems: VTLN, SAT, and the ROVER technique. Applying VTLN yielded improvements of 1.0% on English, 1.7% on Mandarin, and 0.3% on Russian. SAT reduced the WER by 4.5% on Russian and 3.1% on Dari. The ROVER technique yielded improvements in system performance of up to 2.4%. Given the substantial gains in system performance obtained with SAT, recommendations for future work include evaluating SAT across all languages, investigating how much speech data is needed from a single speaker to obtain an improvement in performance, and implementing an automatic method for detecting speaker changes and clustering speakers so that SAT can be applied to data where the speaker boundaries are unknown (*i.e.,* broadcast news).

AF detectors were developed for English using GMMs, two-class MLPs, fusion MLPs, and multi-class MLPs. The outputs of the detectors were used to form feature sets for HMM-based phoneme and word recognizers. On TIMIT, the Fusion-2 feature set yielded an improvement in PER of 3.7% compared to an MFCC system when decoding with monophones. On CSLU, the Fusion-2 features yielded improvements of 2.0% PER compared to MFCCs when decoding with either monophone or triphone models. On SVitchboard, appending the scores from the multi-class MLPs to PLP features yielded an improvement in WER of 6.0%. Finally, appending the scores from the English multi-class MLPs to MFCC features reduced the WER by 1.6% on Russian and 1.4% on Dari. Recommendations for future work include evaluating the English AF detectors across all languages, investigating methods for adapting the multi-class MLPs to different languages, and using alternative acoustic features for input to the MLPs.

Speech synthesis systems were developed for 13 different languages using HTS. Four methods were investigated for modifying these systems: expanding the model set to include additional contextual features, changing the MDL control factor, using speaker recognition scores and/or F0 values for grouping speakers to train voices, and applying speaker adaptation. Two GUIs were also developed for training and evaluating the speech synthesizers. Recommendations for future work include investigating how much speech data is needed to obtain an improvement when using an expanded model set, determining how much speech data is needed for speaker adaptation, and investigating the effects of using different speaker groupings to train the base model that is used for adaptation.

REFERENCES

1. Kong, J. and Graph, D. "TDT4 Multilingual Broadcast News Speech Corpus," *Linguistic Data Consortium*, Philadelphia, 2005. (Available at http://www.ldc.upenn.edu)
2. "CSR-II (WSJ1) Complete," *Linguistic Data Consortium*, Philadelphia, 1994. (Available at http://www.ldc.upenn.edu)
3. Canavan, A. and Zipperlen, G. "CALLHOME Mandarin Chinese Speech," *Linguistic Data Consortium*, Philadelphia, 1996. (Available at http://www.ldc.upenn.edu)
4. "1997 Mandarin Broadcast News Speech (HUB4-NE)," *Linguistic Data Consortium*, Philadelphia, 1998. (Available at http://www.ldc.upenn.edu)
5. Schultz, T. "GlobalPhone: a Multilingual Speech and Text Database Developed at Karlsruhe University," in *Proc. of ICSLP*, Denver, Colorado, 2002.
6. Cambridge University Engineering Department, "The HTK Book," 2007. (Available at http://htk.eng.cam.ac.uk)
7. Clarkson, P. and Rosenfeld, R. "Statistical Language Modeling using the CMU-Cambridge Toolkit," in *Proc. of ESCA Eurospeech*, Rhodes, Greece, September, 1997.
8. Lee, A., Kawahara, T., and Shikano, K. "Julius – an Open Source Real-Time Large Vocabulary Recognition Engine," in *Proc. of EUROSPEECH*, Aalborg, Denmark, September 2001.
9. Welling, L., Ney, H., and Kanthak, S. "Speaker Adaptive Modeling by Vocal Tract Normalization," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 6, September 2002.
10. Fiscus, J. "A Post-Processing System to Yield Reduced Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, December 1997.
11. Pellom, B. *SONIC: The University of Colorado Continuous Speech Recognizer*, Technical Report TR-CSLR-2001-01, University of Colorado, March 2001.
12. International Phonetic Association, "Handbook of the International Phonetic Association," Cambridge University Press, 1999.
13. Reynolds, D., Quatieri, T., and Dunn, R. "Speaker Verification using Adapted Gaussian Mixture Model," *Digital Signal Processing*, Vol. 10, Nos. 1-3, January 2000.
14. Ore, B., and Slyh, R. "Score Fusion for Articulatory Feature Detection," in *Proc. of Interspeech*, Antwerp, Belgium, August 2007.
15. Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. "TIMIT Acoustic Phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, Philadelphia, 1993. (Available at http://www.ldc.upenn.edu)
16. Muthusamy, Y., Cole, R., and Oshika, B. "CSLU: Multi-language Telephone Speech Version 1.2," *Linguistic Data Consortium*, Philadelphia, 1993. (Available at http://www.ldc.upenn.edu)
17. Frankel, J., Magimai-Doss, M., King, S., Livescu, K., and Çetin, Ö., "Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech," in *Proc. of Interspeech*, Antwerp, Belgium, August 2007.
18. Cieri, C., Graff, D., Kimball, O., Miller, D., and Walker, K. "Fisher English Training Part I," *Linguistic Data Consortium*, Philadelphia, 2004. (Available at http://www.ldc.upenn.edu)
19. Cieri, C. Graff, D., Kimball, O., Miller, D., and Walker, K. "Fisher English Training Part

II," *Linguistic Data Consortium*, Philadelphia, 2005. (Available at http://www.ldc.upenn.edu)

20. King, S., Bartels, C., and Bilmes, J. "SVitchboard 1: Small Vocabulary Tasks from SVitchboard 1", in *Proc. of Interspeech*, Libson, Portugal, 2005.
21. Livescu, K., Bezman, A., Borges, N., Yung, L., Çetin, Ö., Frankel, J., King, S., Magimai-Doss, M., Chi, X., and Lavoie, L. "Manual Transcription of Conversational Speech at the Articulatory Feature Level," in *Proc. of ICASSP*, Honolulu, Hawaii, 2007.
22. Kominek, J., and Black, A. "*CMU Arctic Databases for Speech Synthesis*," CMU-LTI-03-177, Language Technologies Institute School of Computer Science, Carnegie Mellon University, 2003. (Available at http://festvox.org/cmu_arctic)
23. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. "The HMM-Based Speech Synthesis System Version 2.0," in *Proc. of ISCA SSW6*, Bonn, Germany, August 2007.
24. Tokuda, K., Mausko, T., and Miyazaki, N. "Multi-Space Probability Distribution HMM," *IEICE Transactions on Information and Systems*, Vol. E85-D, No. 3, March 2002.
25. Shinoda, K., and Watanabe, T. "Acoustic Modeling based on the MDL Principle for Speech Recognition," in *Proc. of Eurospeech*, Rhodes, Greece, September 1997.
26. Garofolo, J., Graff, D., Paul, D., and Pallett, D. "CSR-I (WSJ0) Complete," *Linguistic Data Consortium*, Philadelphia, 2007. (Available at http://www.ldc.upenn.edu)
27. Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. "A Hidden Semi-Markov Model-Based Speech Synthesis System," *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 5, May 2007.
28. Luxburg, U. "*A Tutorial on Spectral Clustering*," Technical Report TR-149, Max Planc Institute for Biological Cybernetics, August 2006.
29. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. "Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 1, January 2009.

LIST OF ACRONYMS & GLOSSARY

AF	Articulatory Feature
AFRL	Air Force Research Laboratory
ARL	Army Research Laboratory
CALLHOME	a speech corpus of unscripted telephone conversations
CMLLR	constrained maximum likelihood linear regression
CMU	Carnegie Mellon University
CSMAPLR	Constrained Structural Aaximum-A-Posteriori Linear Regression
EER	Equal Error Rate
F0	fundamental frequency
Fisher	a speech corpus of telephone conversations
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
GlobalPhone	a multilingual text and speech database
HMM	Hidden Markov Model
HSMM	hidden semi-Markov model
HTK	Hidden Markov ToolKit
HTK	Cambridge University hidden Markov model toolkit
HTS	hidden Markov model based speech synthesis toolkit
HUB4	a broadcast news speech corpus
HDecode	Cambridge University large vocabulary continuous speech recognizer
ICSI	International Computer Science Institute
IPA	International Phonetic Alphabet
Julius	an open source large vocabulary continuous speech recognition engine
KLT	Karhunen Loéve transformation
LASER	Language and Speech exploitation Resources
LM	Language Model
MAP	Maximum-A-Posteriori
MDL	Minimum Description Length
MFCC	Mel-Frequency Cepstral Coefficient
MLP	Multi-Layer Perceptron
MSD	Multi-Space Probability Distribution
PER	Phoneme Error Rate
PLP	Perceptual Linear Prediction
ROVER	Recognizer Output Voting Error Reduction
SAT	Speaker Adaptive Training
SI	Speaker Independent
SONIC	University of Colorado continuous speech recognizer
SPTK	Speech Signal Processing Toolkit
SRover	University of Brno implementation of recognizer output voting error reduction
TDT4	phase four of the topic detection and tracking corpus
TRANSTAC	Translation System for Tactical Use
VTLN	Vocal Tract Length Normalization

WER	Word Error Rate
WSJ0	phase one of the Wall Street Journal corpus
WSJ1	phase two of the Wall Street Journal corpus